

# 2017 Machine Learning with R

## Association Rules

강필성

고려대학교 산업경영공학부

pilsung\_kang@korea.ac.kr

# 목차

I

연관규칙분석:A Priori Algorithm

II

연관규칙분석: 활용 분야

III

R 실습

# 한번쯤은 봤을 법한...



크게 보기 | 미리 보기

영업점 재고 · 위치 >

Klover 평점 102명

안내

오늘의책 북모닝 CEO 이벤트 무료배송 사은품

## 호모 데우스 미래의 역사

유발 하라리 지음 | 김명주 옮김 | 김영사 | 2017년 05월 19일 출간

★★★★★ 리뷰 30개 [리뷰쓰기](#)

국내도서 주간베스트 8위 | 인문 주간베스트 2위

주요 일간지 북섹션 추천도서 ▾

정가 : 22,000원

판매가 : **19,800원** [10%↓ 2,200원 할인]

제휴할인가 : **18,810원** 교보-KB국민카드 5% 청구할인(실적무관) [카드/포인트 안내](#)

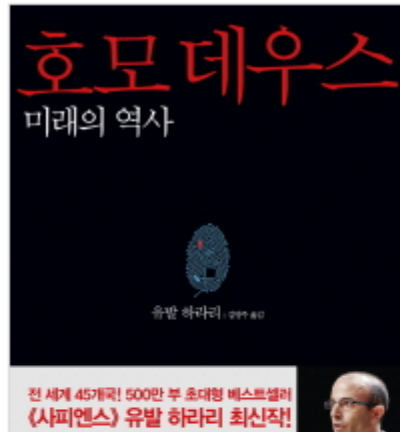
통합포인트 : **1,100원** 적립 [5% 적립]

추가혜택 : **N Pay** 네이버페이 결제 시 최대 **2%** 추가 적립 >

**PAYCO** 페이코 결제 시 **6,500원** 할인 + **1만원** 적립 >

**L.POINT** 실 결제 금액의 **0.5%** 적립 [안내](#)

# 한번쯤은 봤을 법한...



오늘의책 북모닝CEO 이벤트 무료배송 사은품

## 호모 데우스 미래의 역사

유발 하라리 지음 | 김명주 옮김 | 김영사 | 2017년 05월 19일 출간

★★★★★ 리뷰 30개 [리뷰쓰기](#)

국내도서 주간베스트 8위 | 인문 주간베스트 2위

주요 일간지 북섹션 추천도서 ▼

정가 : 22,000원

이 책을 구매한 분들이 함께 구매한 상품입니다

전체선택

장바구니 담기



말의 품격

13,050원



라틴어 수업(지적이고 아름다운 삶을 위한)

13,500원



인간의 위대한 여정(양장본 HardCover)

19,800원



사피엔스

19,800원



이탈리아의 사생활

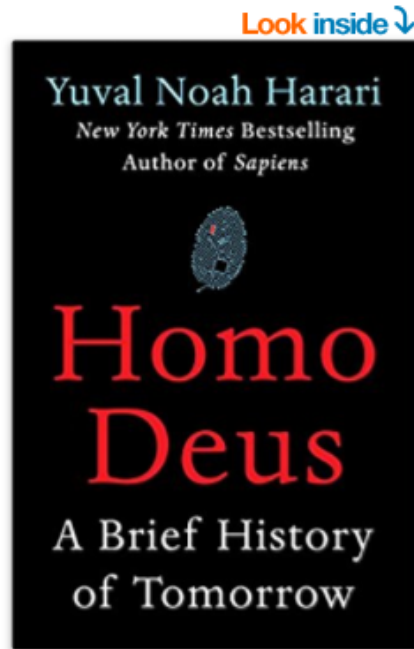
14,400원



한식의 품격(반양장)

16,200원

# 한번쯤은 봤을 법한...



 Listen



[See this image](#)

## Homo Deus: A Brief History of Tomorrow Hardcover – February 21, 2017

by Yuval Noah Harari (Author)

★★★★☆ 444 customer reviews

[Amazon Charts](#) #16 Most Read

[See all 14 formats and editions](#)

Kindle  
\$15.50


Read with Our **Free App**

**Hardcover**  
**\$19.88**

12 Used from \$15.00  
51 New from \$15.25  
1 Collectible from \$84.95

Paperback  
from \$8.99

15 Used from \$9.48  
34 New from \$8.99

 Audiobook  
\$0.00

Free with your Audible trial

Audio CD  
\$30.29

5 Used from \$24.17  
27 New from \$21.57

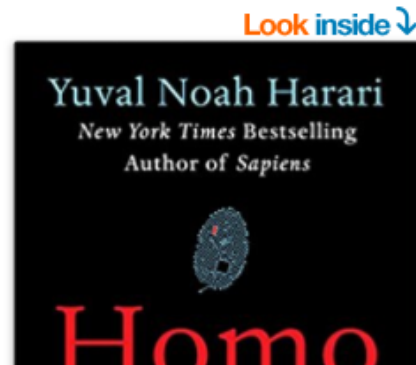
### NEW YORK TIMES BESTSELLER

Yuval Noah Harari, author of the critically-acclaimed *New York Times* bestseller and international phenomenon *Sapiens*, returns with an equally original, compelling, and provocative book, turning his focus toward humanity's future, and our quest to upgrade humans into gods.

Over the past century humankind has managed to do the impossible and rein in famine, plague, and war. This may seem hard to accept, but, as Harari explains in his trademark style—thorough, yet riveting—famine, plague and war have been transformed from incomprehensible and uncontrollable forces of nature into manageable challenges. For the first time ever, more people die from eating too much than

[Read more](#)

# 한번쯤은 봤을 법한...



Look inside ↴

## Homo Deus: A Brief History of Tomorrow Hardcover – February 21, 2017

by Yuval Noah Harari ▾ (Author)

★★★★☆ ▾ 444 customer reviews

Amazon Charts #16 Most Read

▶ See all 14 formats and editions

Kindle  
\$15.50

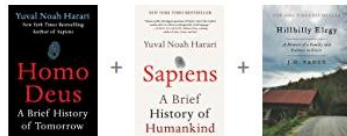
Hardcover  
**\$19.88**

Paperback  
from \$8.99

Audiobook  
\$0.00

Audio CD  
\$30.29

Frequently bought together



Total price: **\$57.67**

Add all three to Cart

Add all three to List

✓ **This item:** Homo Deus: A Brief History of Tomorrow by Yuval Noah Harari Hardcover **\$19.88**

✓ Sapiens: A Brief History of Humankind by Yuval Noah Harari Hardcover **\$21.00**

✓ Hillbilly Elegy: A Memoir of a Family and Culture in Crisis by J. D. Vance Hardcover **\$16.79**

Customers who bought this item also bought

Page 1 of 13

Sapiens: A Brief History of Humankind  
Yuval Noah Harari  
★★★★☆ 3,286  
Hardcover  
**\$21.00** ✓prime

A Full Life: Reflections at Ninety  
Jimmy Carter  
★★★★☆ 261  
Hardcover  
64 offers from **\$7.15**

The Inevitable: Understanding the 12 Technological Forces That Will Shape Our Future  
Kevin Kelly  
★★★★☆ 225  
Paperback  
**\$12.18** ✓prime

The Heart: A Novel  
Maylis de Kerangal  
★★★★☆ 82  
Hardcover  
**\$19.93** ✓prime

The Heart: A Novel  
Maylis de Kerangal  
★★★★☆ 82  
Paperback  
**\$11.00** ✓prime

Hillbilly Elegy: A Memoir of a Family and Culture in Crisis  
J. D. Vance  
★★★★☆ 7,255  
Hardcover  
**\$16.79** ✓prime

Summary: Sapiens: A Brief History of Humankind  
Yuval Noah Harari  
★★★★☆ 14  
Paperback  
**\$9.99** ✓prime

Easternization: Asia's Rise and America's Decline From Obama to Trump...  
Gideon Rachman  
★★★★☆ 12  
Hardcover  
**\$17.07** ✓prime

# 일상 생활에서의 연관 규칙 분석

## ❖ 장바구니 분석: Market basket analysis (MBA)



Wall Mart (USA)



E-Mart (Korea)

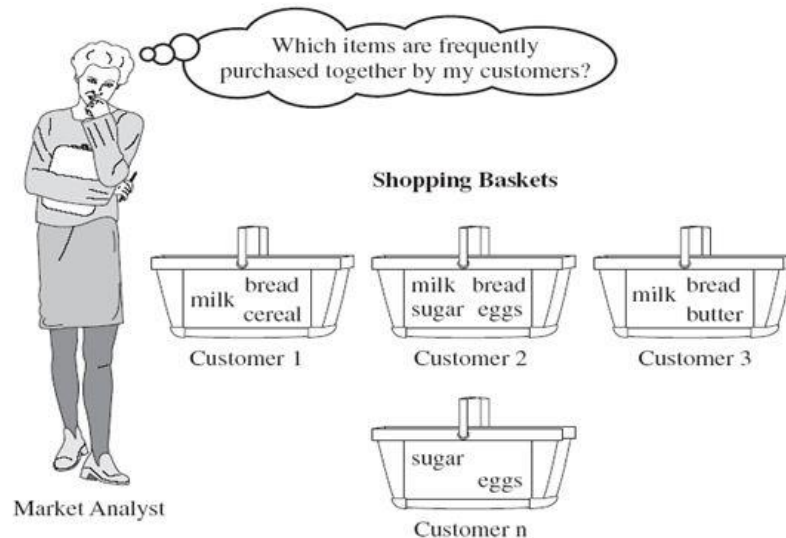




# 연관규칙 분석

## ❖ 목적

- 어떤 두 아이템 집합이 **빈번히 발생**하는가를 알려주는 **일련의 규칙**들을 생성
  - ✓ Produce rules that define “what goes with what”
- 우리의 데이터에 의하면 “X 아이템을 구매하는 고객들은 Y 아이템 역시 구매할 가능성이 높다” → 콘텐츠 기반 추천 시스템에 널리 사용
- 장바구니 분석(Market Basket Analysis)으로도 널리 알려짐





# 연관규칙 분석

## ❖ 데이터 속성



- 각 레코드는 트랜잭션의 형태를 가짐
- 행렬의 형태로 표현하게 되면 대부분의 셀이 0의 값은 갖는 희소행렬(sparse matrix)이 됨

Tid	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Tid	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

# 연관규칙 분석: 예제

## ❖ 동네 작은 가게 매출 트랜잭션 데이터

Transaction	Item 1	Item 2	Item 3	Item 4
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

# 연관규칙 분석: 용어 및 규칙 생성

## ❖ 용어: Terminology

- 조건절(Antecedent) – “IF” part
- 결과절(Consequent) – “THEN” part
- 아이템 집합(Item set) – 조건절 또는 결과절을 구성하는 아이템들의 집합
- 조건절 아이템 집합과 결과절 아이템 집합은 상호배반 (한 아이템이 조건절과 결과절에 모두 포함될 수 없음)

## ❖ 규칙 생성: Generating rules

- 매우 많은 수의 규칙이 생성 가능 (예시: 첫번째 트랜잭션)
  - ✓ 계란을 구매하는 사람들은 라면도 함께 구매한다.
  - ✓ 계란과 라면을 구매하는 사람들은 참치도 함께 구매한다.
  - ✓ 참치를 구매하는 사람들은 계란도 함께 구매한다.
  - ✓ ...

# 연관규칙 분석: 규칙의 효용성 측정 지표

For the rule  $A \rightarrow B$

❖ 지지도: Support

$$\text{support}(A) = P(A)$$

- 빈발 아이템 집합(frequent item sets)을 판별하는데 사용

❖ 신뢰도: Confidence

$$\text{confidence}(A \rightarrow B) = \frac{P(A, B)}{P(A)}$$

- 아이템 집합 간의 연관성 강도를 측정하는데 사용

❖ 향상도: Lift

$$\text{lift}(A \rightarrow B) = \frac{P(A, B)}{P(A) \cdot P(B)}$$

- 생성된 규칙이 실제 효용가치가 있는지를 판별하는데 사용

# 연관규칙 분석: 규칙의 효용성 측정 지표

Rule:  $X \Rightarrow Y$

$$Support = \frac{freq(X, Y)}{N}$$

$$Confidence = \frac{freq(X, Y)}{freq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$



Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

# 연관규칙 분석: 규칙 생성

## ❖ 유용한 연관 규칙들을 어떻게 찾아낼 것인가?

- 이상적으로는 모든 생성 가능한 규칙을 만든 뒤, 각 규칙의 지지도, 신뢰도, 향상도를 측정하여 유용한 규칙들만을 찾아냄
- 아이템 수가 증가할수록 계산에 소요되는 시간이 기하급수적으로 증가함

## ❖ Brute-force approach

- 가능한 모든 규칙을 나열함
- 모든 규칙의 지지도와 신뢰도를 계산함
- 최소지지도와 최소신뢰도 조건을 만족하지 못하는 규칙을 제거
- Computationally prohibitive!

# 연관규칙 분석: 규칙 생성

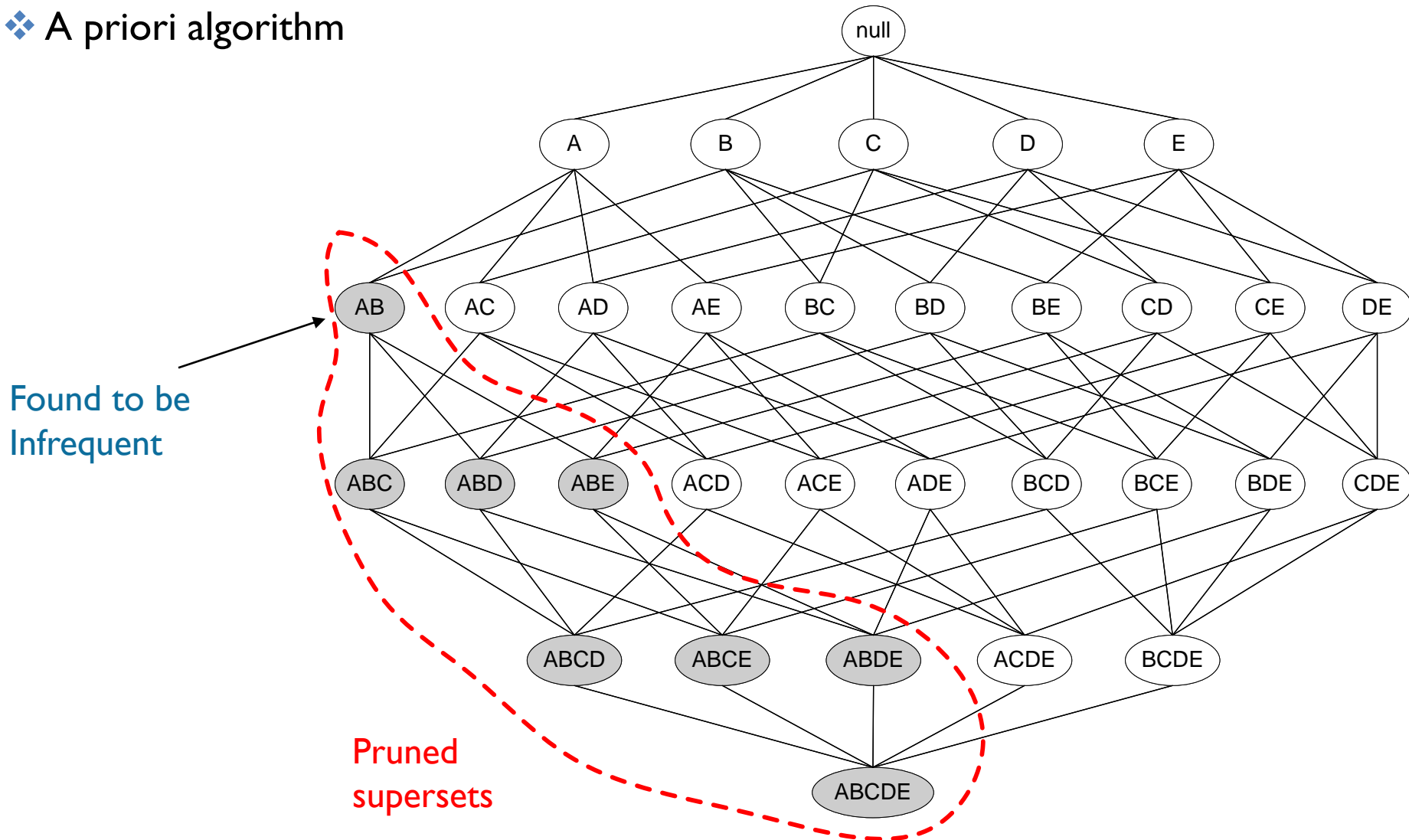
## ❖ A priori algorithm

- 빈발 집합(frequent item sets)만을 고려하여 규칙 생성
- **지지도(support)**
  - ✓ 조건절에 속하는 아이템 집합이 발생할 확률
  - ✓ 아이템 집합 {계란, 라면}의 지지도는 40%
- **최소 지지도(minimum support)**
  - ✓ 유용한 규칙으로 인정받기 위해 필요한 최소 지지도
- 최소 지지를 만족하지 못하는 아이템 집합의 상위집합(superset)은 항상 최소 지지를 만족하지 않음
  - ✓ Support of an item set never exceeds the support of its subsets, which is known as **anti-monotone** property of support.



# 연관규칙 분석: 규칙 생성

## ❖ A priori algorithm



# 연관규칙 분석: 빈발 아이템 집합 생성

## 최소 지지도 조건 부여

- 최소 지지도: 2 transactions or 20%

Transaction	Item 1	Item 2	Item 3	Item 4
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

# 연관규칙 분석: 빈발 아이템 집합 생성

2

최소 지지도 조건을 만족하는 1개짜리 아이템 집합을 생성

- $\text{Support \{noodle\}} = 8/10 = 80\%$
- $\text{Support \{egg\}} = 5/10 = 50\%$
- $\text{Support \{cola\}} = 5/10 = 50\%$
- $\text{Support \{rice\}} = 3/10 = 30\%$
- $\text{Support \{tuna\}} = 2/10 = 20\%$
- $\text{Support \{onion\}} = 1/10 = 10\%$

양파(onion)는 최소 지지도 조건을 만족하지 못했으므로 이후 분석에서 제외

# 연관규칙 분석: 빈발 아이템 집합 생성

3

앞 단계에서 살아남은 아이템들을 이용하여 최소 지지도 조건을 만족하는 2개짜리 아이템 집합을 생성

	noodle	egg	cola	rice	tuna
noodle		40%	40%	20%	20%
egg			30%	0%	20%
cola				0%	10%
rice					0%
tuna					

- {noodle, egg}, {noodle, cola}, {noodle, rice}, {noodle, tuna}, {egg, cola}, {egg, tuna} are frequent two-item sets.

# 연관규칙 분석: 빈발 아이템 집합 생성

더 이상 최소 지지도 이상을 나타내는 아이템 집합이 없을 때까지 아이템 집합의 크기를 1씩 증가시키면서 반복 수행

4

Set-size	Item 1	Item 2	Item 3	...	Item 6
1	noodle				
1	egg				
1	cola				
1	rice				
1	tuna				
2	noodle	egg			
2	noodle	cola			
2	noodle	rice			
...	...	...			

# 연관규칙 분석: A Priori Algorithm

## ❖ A priori algorithm

- Let  $k=1$
- Generate frequent itemsets of length 1
- Repeat until no new frequent itemsets are identified
  - ✓ Generate length  $(k+1)$  candidate itemsets from length  $k$  frequent itemsets
  - ✓ Prune candidate itemsets containing subsets of length  $k$  that are infrequent
  - ✓ Count the support of each candidate by scanning the DB
  - ✓ Eliminate candidates that are infrequent, leaving only those that are frequent

# 연관규칙 분석: 규칙 평가

## ❖ 신뢰도: Confidence

- 조건절이 발생했다는 가정 하에 결과절이 발생할 조건부 확률
- E.g. “if noodle is purchased, then egg is also purchased”

$$\text{support}(\text{noodle}) = P(\text{noodle}) = \frac{8}{10}, \quad \text{support}(\text{egg}) = P(\text{egg}) = \frac{5}{10}$$

$$\text{confidence}(\text{noodle} \rightarrow \text{egg}) = \frac{P(\text{noodle}, \text{egg})}{P(\text{noodle})} = \frac{4/10}{8/10} = 0.5(50\%)$$

- 비교 대상 신뢰도(benchmark confidence): 전체 트랜잭션에서 결과절이 발생할 확률 ( $P(\text{egg})$ ,  $\text{support}(\text{egg})$ )
- 규칙 ( $\text{noodle} \rightarrow \text{egg}$ )의 신뢰도가  $P(\text{egg})$ 보다 작으면 규칙으로서의 효용 가치는 낮음



# 연관규칙 분석: 규칙 평가

## ❖ 지지도: Lift

- 신뢰도/비교 대상 신뢰도: Confidence/(benchmark confidence)

$lift(noodle \rightarrow egg)$

$$\begin{aligned}
 &= \frac{confidence(noodle \rightarrow egg)}{support(egg)} = \frac{P(noodle, egg)}{P(egg)} = \frac{P(noodle, egg)}{P(noodle) \times P(egg)} \\
 &= \frac{\frac{4}{10}}{\frac{8}{10} \times \frac{5}{10}} = 1
 \end{aligned}$$

- 신뢰도 = 1일 경우, 조건절과 결과절은 통계적으로 독립사건임을 의미함 → 규칙 사이에 유의미한 연관성이 없음
- 신뢰도 > 1일 경우 조건절과 결과절은 서로 긍정적인 연관관계를 나타냄

# 연관규칙 분석: 사례 결과

## ❖ 규칙 생성을 위한 기준 지지도 및 신뢰도 설정

- 기준 지지도: 20%
- 기준 신뢰도: 70%

Rule #	Antecedent (a)	Consequent	Support	Confidence	Lift
1	tuna=>	egg, noodle	2	100	2.5
2	tuna=>	egg	2	100	2
3	noodle, tuna=>	egg	2	100	2
4	rice=>	noodle	3	100	1.25
5	egg, tuna=>	noodle	2	100	1.25
6	tuna=>	noodle	2	100	1.25
7	cola=>	noodle	5	80	1
8	egg=>	noodle	5	80	1

# 연관규칙 분석: 요약

## ❖ 연관규칙분석

- 트랜잭션 데이터베이스에 존재하는 아이템 집합들 간의 연관성을 나타내는 규칙을 생성하는 분석 기법
- 다양한 분야의 추천 시스템 구축에 널리 사용됨
- 전체 규칙을 모두 생성하는 것이 비효율적이기 때문에 효율적인 빈발 집합을 찾아내는 A Priori 알고리즘을 사용
- 규칙의 효용성은 지지도, 신뢰도, 향상도의 세 가지를 이용하여 평가
- 규칙 1:  $A \rightarrow B$ 와 규칙 2:  $C \rightarrow D$ 에 대해 지지도, 신뢰도, 향상도가 모두 클 경우에만 규칙 1이 규칙 2보다 효과적인 규칙으로 결론지을 수 있음

# 목차

I

연관규칙분석:A Priori Algorithm

II

연관규칙분석: 활용 분야

III

R 실습

# 사례 1: 유통에서의 연관규칙분석

## ❖ 국내 실제 유통회사 데이터를 이용한 분석 사례

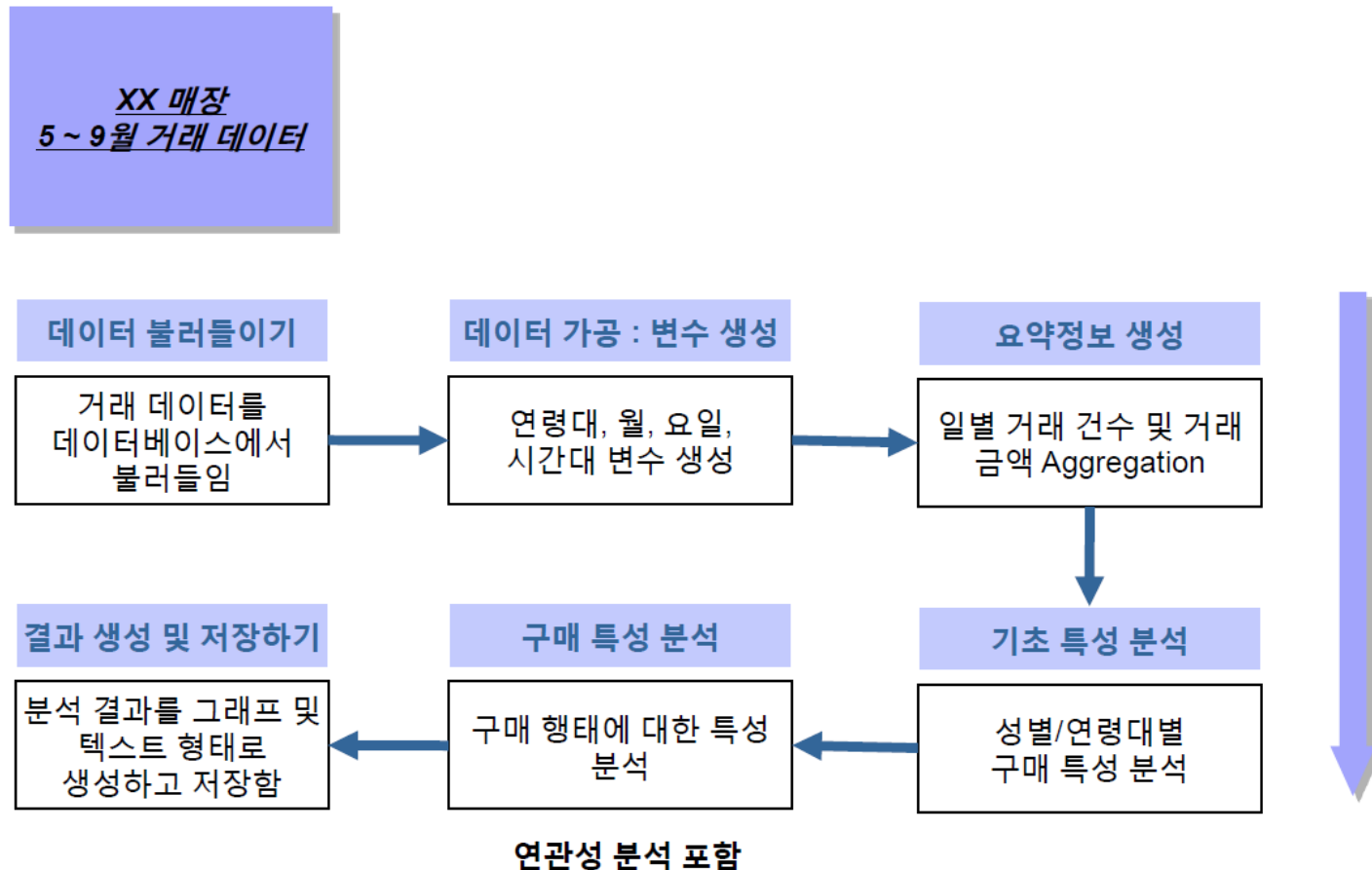
### ■ 데이터 구조

변수명	설명
회원번호	개인 고유 식별 번호
성별	남자 or 여자
연령대	회원의 생년월일을 기준으로 연령대 구분
발급일자	회원 가입일자
매출일자	상품 구매 일자
매출요일	상품 구매 요일
평일공휴일구분	평일 혹은 공휴일 구분 (공휴일은 토요일, 일요일, 휴일)
대분류명	상품의 대분류 구분
중분류명	상품의 중분류 구분
소분류명	상품의 소분류 구분
세분류명	상품의 세분류 구분
수량	상품 구매 개수
매출금액	상품 구매 금액

# 사례 1: 유통에서의 연관규칙분석

## ❖ 국내 실제 유통회사 데이터를 이용한 분석 사례

### ■ 데이터 분석 프로세스



# 사례 1: 유통에서의 연관규칙분석

## ❖ 국내 실제 유통회사 데이터를 이용한 분석 사례

### ■ 데이터 분석 시나리오

	<u>시나리오 명</u>	<u>분석 방향</u>
시나리오 1	기초 특성 분석	<ul style="list-style-type: none"> <li>• 상세 분석을 위한 데이터 파악</li> <li>• 구매 현황 리포트 작성</li> </ul>
시나리오 2	그룹별 구매 패턴 분석	<ul style="list-style-type: none"> <li>• 전체, 성별/연령대별 상품 구매 패턴 파악</li> </ul>
시나리오 3	개인별 구매 패턴 분석	<ul style="list-style-type: none"> <li>• 개인별 상품 구매 패턴 파악</li> </ul>
시나리오 4	연관성 분석	<ul style="list-style-type: none"> <li>• 주로 구매하는 상품으로 장바구니를 구성함</li> </ul>



## ■ 구매 물품의 시각화

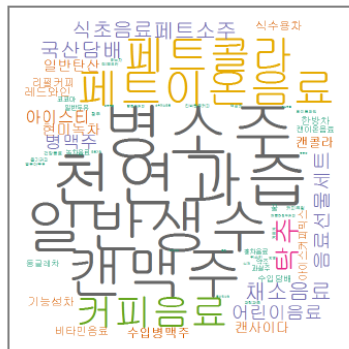


# 사례 1: 유통에서의 연관규칙분석

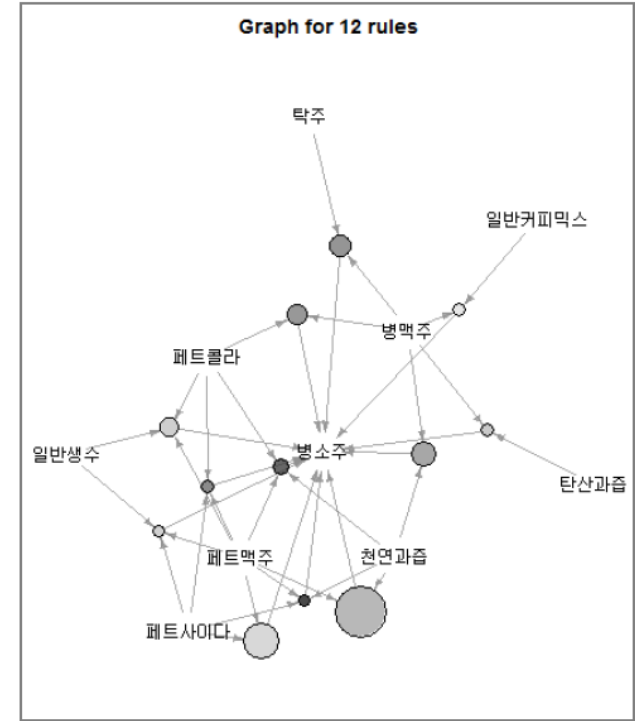
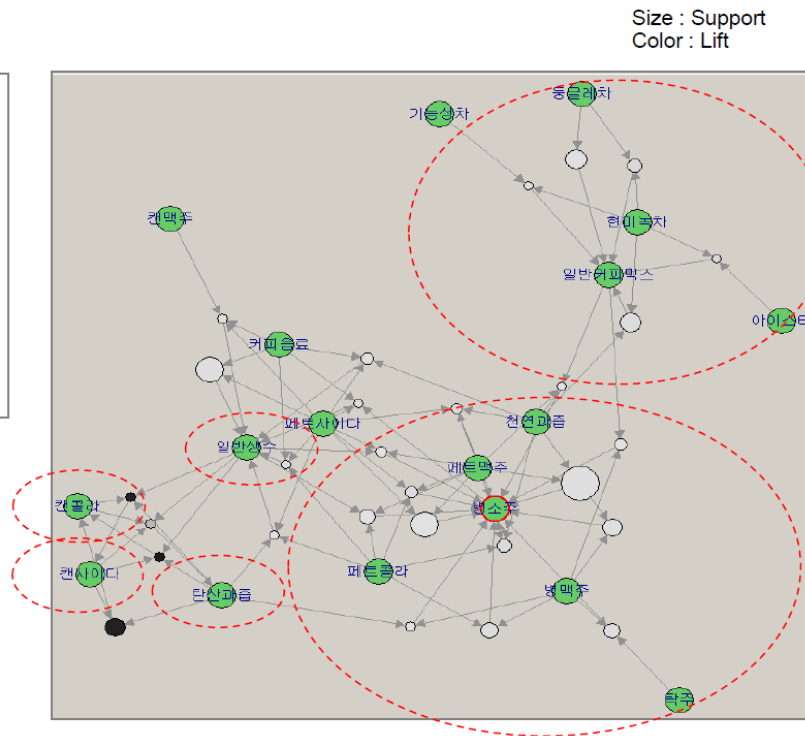
## ❖ 국내 실제 유통회사 데이터를 이용한 분석 사례

### ■ 연관규칙 분석을 통한 상품집합 도출

#### ● 주류,기호



캔콜라	2
캔사이다	1
탄산과즙	1
일반커피믹스	5
일반생수	6
병소주	13



# 사례 2: 동영상 클립 추천

## ❖ 광고 수익 증대를 위한 동영상 추천 시스템 구축

### 하루 평균 1억 건 이상의 데이터를 처리

동영상·광고의 로그 및 메타정보를 raw데이터로 저장·관리하며  
수집 항목 및 사이트는 꾸준히 증가하고 있음

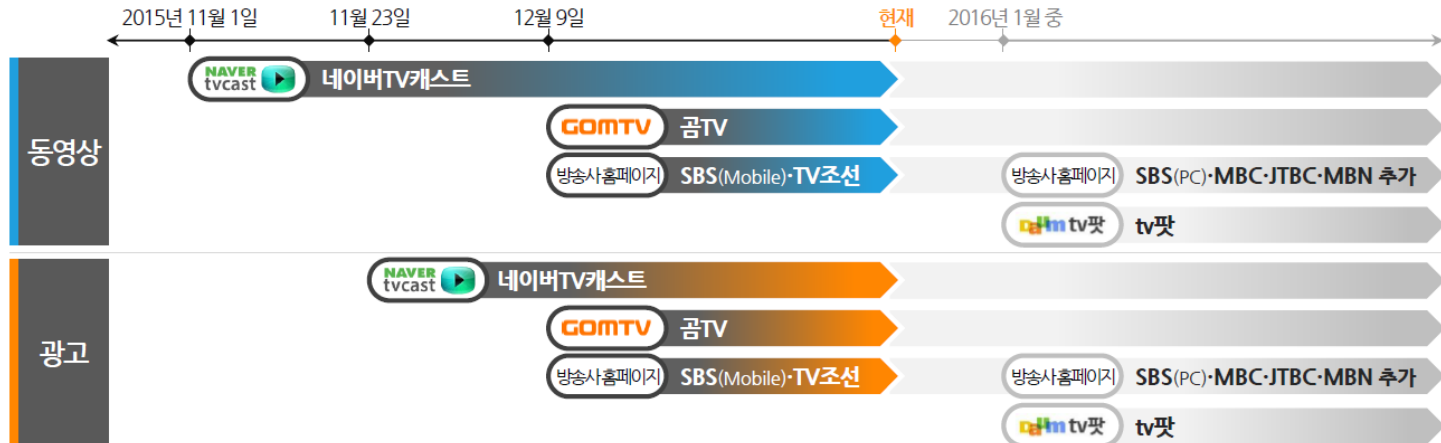
#### ■ 하루 평균 처리하는 Data량 (15년 12월 기준)

**1억 건(30억 건/월)**

- ※ 월 평균 광고노출수 5억 \* 광고로그 5개 + 동영상 조회수 5억 \* 동영상로그 1개 + α(메타정보)
- ※ 현재 동영상로그가 1개이지만, 협의를 통해 이용시간 확보를 위해 5개로 늘릴 예정임
- ※ 데이터 신뢰성, 타당도 높음

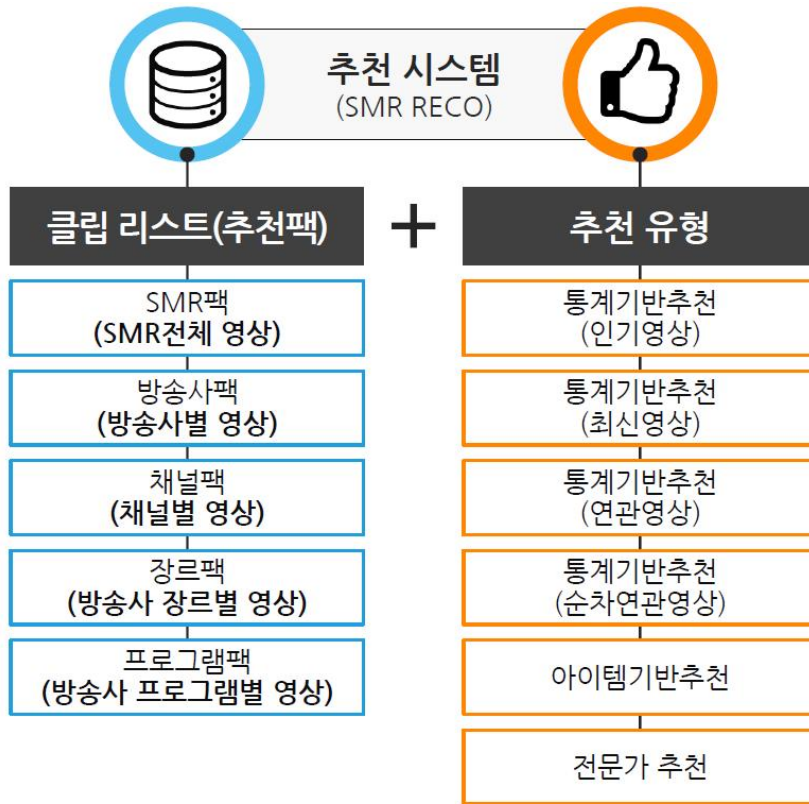
더보기

#### ■ 데이터 수집 사이트 현황 (15년 12월 기준)



# 사례 2: 동영상 클립 추천

## ❖ 광고 수익 증대를 위한 동영상 추천 시스템 구축



## 사례 2: 동영상 클립 추천

### ❖ 광고 수익 증대를 위한 동영상 추천 시스템 구축

유희열의 스케치북 + 구독 17,185 방송 다시보기 구독 66,595명

음악이 필요한 순간 Melon 야생화 박효신  
작사·박효신, 김지향 / 작곡·박효신, 정재일

▶ 1,336,050 | 등록 2016.10.30. | 3,152 | ^ 접기

17,428

# 사례 2: 동영상 클립 추천

## ❖ 광고 수익 증대를 위한 동영상 추천 시스템 구축

### 함께 재생된 동영상



손승연, 전율의 괴물보컬 '금지된 사랑'  
 불후의 명곡2 전설을 노래하다  
 > 243,551 ♡ 825



이승철&정인, EDM으로 재탄생 '서쪽 하늘'  
 불후의 명곡2 전설을 노래하다  
 > 226,822 ♡ 838



첫눈처럼 너에게 가겠다 - 에일리  
 유희열의 스케치북  
 > 711,441 ♡ 5,741



손승연, 파워풀한 가창력+애절한 '사랑 안 해'  
 불후의 명곡2 전설을 노래하다  
 > 266,040 ♡ 1,056



에일리, 압도적인 가창력 '편지'  
 불후의 명곡2 전설을 노래하다  
 > 363,955 ♡ 1,956



황치열&정리, '이브의 경고' 커플 댄스  
 불후의 명곡2 전설을 노래하다  
 > 24,410 ♡ 751



치열이의 치열한 신고식~! '복고 댄스'  
 불후의 명곡2 전설을 노래하다  
 > 9,623 ♡ 365



[1회] 지리산 소년 김영근 - "Lay Me Down"  
 슈퍼스타K 2016  
 > 1,965,325 ♡ 9,368



마마무, 센스 있는 개사가 돋보인 축하무대 '데칼코마니'  
 청룡영화상  
 > 2,795,060 ♡ 21,404



[1회] 한 곡으로는 모자라! 김영근 - "탈진"  
 슈퍼스타K 2016  
 > 1,897,812 ♡ 8,224



'아기해마' VS '다이빙 소년'의 1라운드 무대! - Sea of Love  
 복면가왕  
 > 88,800 ♡ 499



한동근&최효인, '거짓말 거짓말 거짓말' 역대급 소름 무대!~  
 듀엣가요제  
 > 494,583 ♡ 2,651



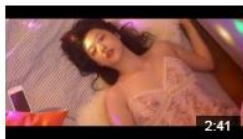
사랑스러운 목소리의 '올리브스녀'는 에이핑크 오하영!  
 복면가왕  
 > 139,772 ♡ 1,581



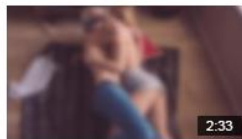
'불광동 휘발유'의 3라운드 무대! - 헤어지는 중입니다  
 복면가왕  
 > 555,697 ♡ 3,786



[15화 예고] 김고은, 공유 향한 작별인사?! '또 만나요.'  
 tvN 10주년 특별기획 <도깨비>  
 > 1,970,210 ♡ 3,149



영화 '리얼' 설리(최진리) 전라신 편집 없이 개봉 결정  
 뉴스릴처 연예TV  
 > 1,454,008 ♡ 937



영화 '리얼' 설리, 전라 노출 상영 인포케이  
 > 427,126 ♡ 321



[풀버전] 주노플로 @ 2차 예선 full ver.  
 소미더머니6  
 > 272,308 ♡ 2,603



황광희 X 개코 - 당신의 밤 (feat.오혁)  
 무한도전  
 > 2,386,613 ♡ 22,395



'올리브스녀' VS '뽀빠이'의 1라운드 무대! - 숲소리  
 복면가왕  
 > 69,793 ♡ 668



[유립피안 챔피언스] 이보다 더 완벽할 순 없다. 딕 야스퍼스  
 당구전문 인터넷 방송국 코즘코리아  
 > 15,197 ♡ 8



- 연구 목적: 근접무선통신기술(NFC)을 이용하여 수집한 박람회 관람객 행동 데이터를 분석하여 전시부스간 연관관계 분석 및 부스 배치 최적화
- 데이터 수집 체계





# 사례 3: 박람회 부스 이용 패턴 분석

## ❖ “2013 내나라 여행 박람회” 데이터를 이용한 분석 사례

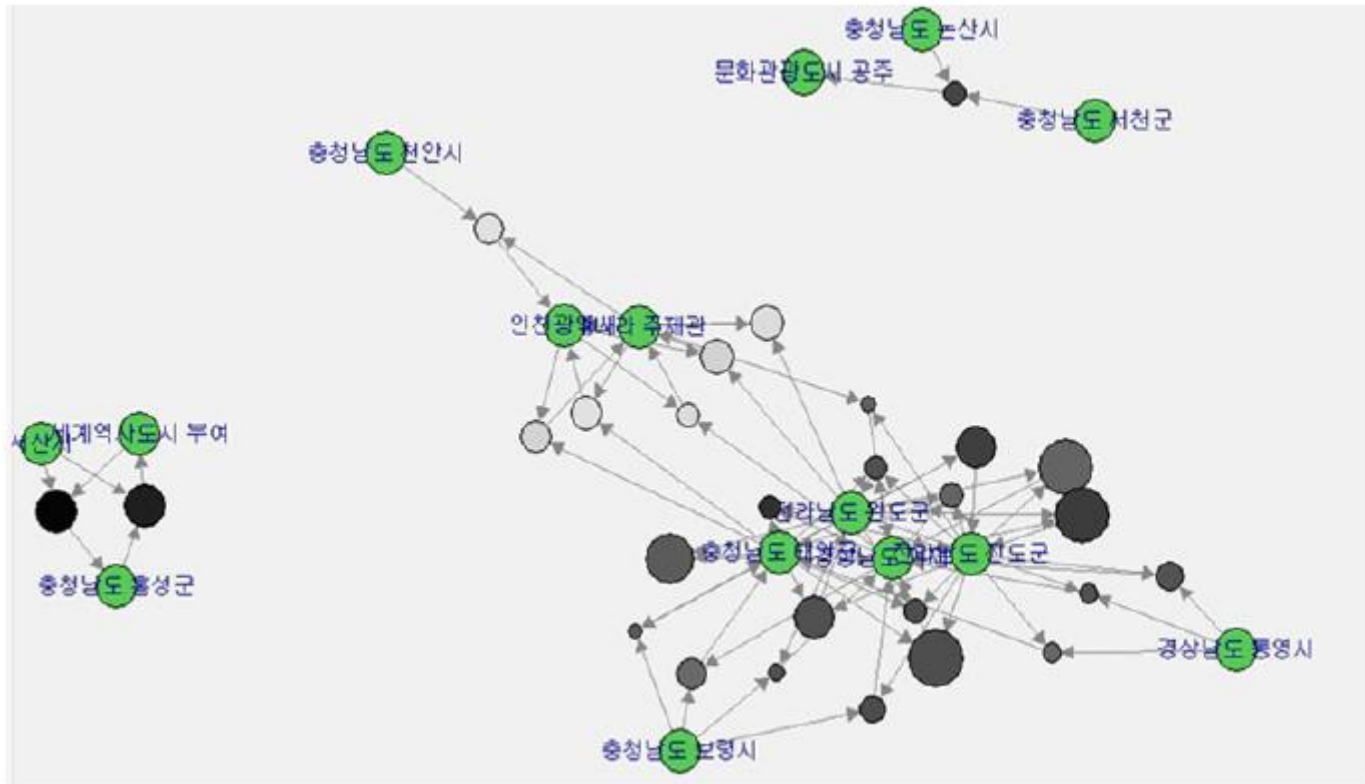
### ■ 연관규칙 분석 결과

	rules	support	confidence	lift
1	{충청남도 논산시, 충청남도 서천군} => {문화관광도시 공주}	0.09631728	0.8500000	4.148157
2	{서산시, 세계역사도시 부여} => {충청남도 홍성군}	0.10859301	0.9274194	5.280307
3	{서산시, 충청남도 홍성군} => {세계역사도시 부여}	0.10859301	0.8156028	4.798463
4	{전라남도 진도군, 충청남도 보령시} => {경상남도 거제시}	0.09820585	0.8888889	4.040057
5	{전라남도 완도군, 충청남도 보령시} => {경상남도 거제시}	0.09254013	0.8750000	3.976931
6	{전라남도 완도군, 충청남도 보령시} => {충청남도 태안군}	0.09065156	0.8571429	3.830018
7	{경상남도 거제시, 충청남도 보령시} => {충청남도 태안군}	0.10103872	0.8045113	3.594842
8	{경상남도 통영시, 전라남도 진도군} => {경상남도 거제시}	0.10009443	0.8688525	3.948990
9	{경상남도 통영시, 전라남도 진도군} => {충청남도 태안군}	0.09348442	0.8114754	3.625960
10	{경상남도 통영시, 전라남도 완도군} => {경상남도 거제시}	0.09348442	0.8761062	3.981959
11	{전라남도 완도군, 전라남도 진도군} => {경상남도 거제시}	0.11709160	0.8104575	3.683582
12	{경상남도 거제시, 전라남도 완도군} => {전라남도 진도군}	0.11709160	0.8157895	4.319605
13	{전라남도 진도군, 충청남도 태안군} => {전라남도 완도군}	0.10859301	0.8156028	3.980292
14	{전라남도 완도군, 충청남도 태안군} => {전라남도 진도군}	0.10859301	0.8041958	4.258217
15	{전라남도 진도군, 충청남도 태안군} => {경상남도 거제시}	0.11803588	0.8865248	4.029312
16	{내나라 주재관, 전라남도 진도군} => {경상남도 거제시}	0.09159585	0.8151261	3.704800
17	{전라남도 완도군, 충청남도 태안군} => {경상남도 거제시}	0.11331445	0.8391608	3.814040
18	{내나라 주재관, 전라남도 완도군} => {인천광역시}	0.10387158	0.8396947	1.967338
19	{인천광역시, 전라남도 완도군} => {내나라 주재관}	0.10387158	0.8661417	2.084646
20	{내나라 주재관, 충청남도 천안시} => {인천광역시}	0.10103872	0.8045113	1.884906
21	{경상남도 거제시, 인천광역시} => {내나라 주재관}	0.09631728	0.8225806	1.979802
22	{내나라 주재관, 충청남도 태안군} => {인천광역시}	0.10198300	0.8059701	1.888324
23	{인천광역시, 충청남도 태안군} => {내나라 주재관}	0.10198300	0.8571429	2.062987
24	{경상남도 거제시, 전라남도 완도군, 전라남도 진도군} => {충청남도 태안군}	0.09631728	0.8225806	3.675582
25	{전라남도 완도군, 전라남도 진도군, 충청남도 태안군} => {경상남도 거제시}	0.09631728	0.8869565	4.031274
26	{경상남도 거제시, 전라남도 진도군, 충청남도 태안군} => {전라남도 완도군}	0.09631728	0.8160000	3.982230
27	{경상남도 거제시, 전라남도 완도군, 충청남도 태안군} => {전라남도 진도군}	0.09631728	0.8500000	4.500750

## 사례 3: 박람회 부스 이용 패턴 분석

## ❖ “2013 내나라 여행 박람회” 데이터를 이용한 분석 사례

## ■ 연관규칙 분석 결과

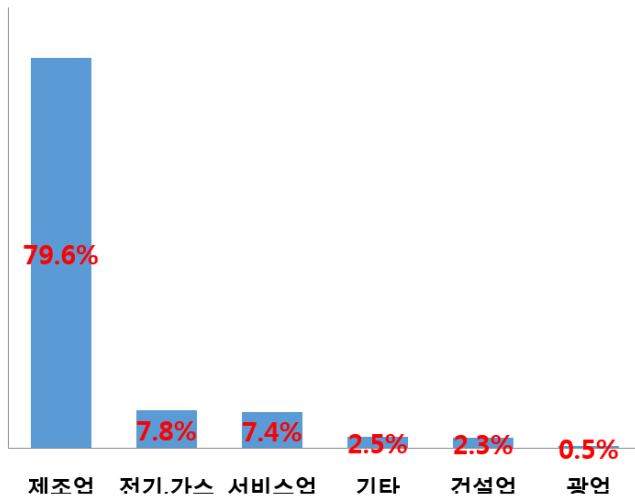


# 사례 4: 기업체 교육프로그램 추천

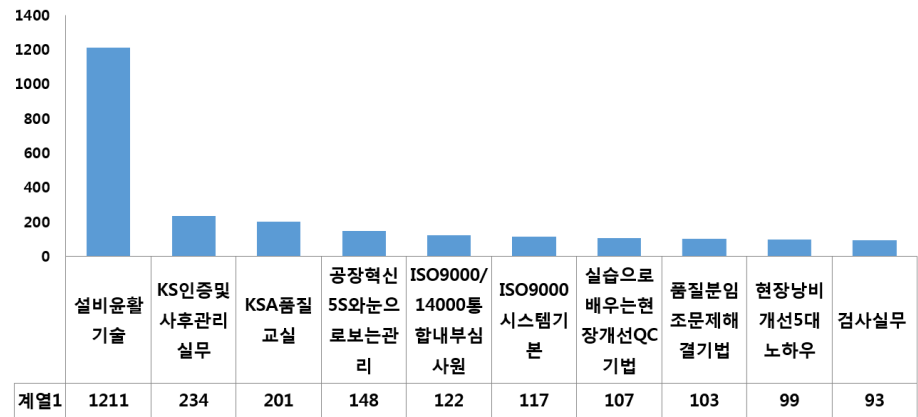
## ❖ 국내 K사의 기업체 교육 프로그램 운영 현황

- 교육 대상 기업의 주 업종은 제조업으로써 전체 교육 이수 기업의 79.6%에 해당함
- 대부분의 기업이 1~5개의 교육을 이수하고 있음

교육기관	기간	참여기업(수)	교육프로그램(수)	참가횟수
K사	2012 년 1월~8월	3,604	263	7,484



<참여기업의 업종별 분류>



<참여기업수 상위 10개 프로그램>

## 사례 4: 기업체 교육프로그램 추천

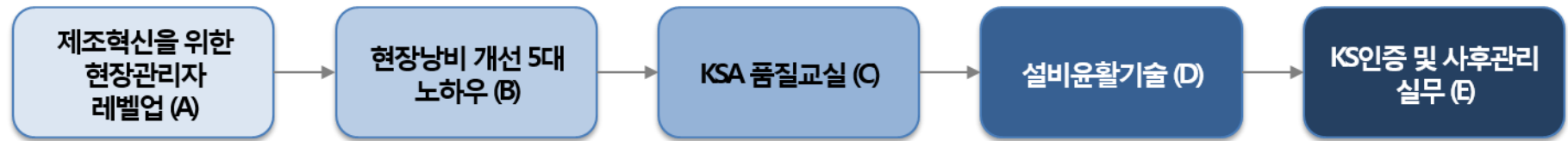
### ❖ 연관규칙분석 수행

- 지지도와 신뢰도의 기준값을 변화시켜 가며 총 25가지의 조합에 대해 생성된 규칙을 조사
- 현실적으로 활용 가능한 20~40개의 규칙을 생성한 세 가지 조합에 대한 추가 분석 실시
  - ✓ Case 1: 지지도 0.01, 신뢰도 0.2, Case 2: 지지도 0.01, 신뢰도 0.2, Case 3: 지지도 0.005, 신뢰도 0.3

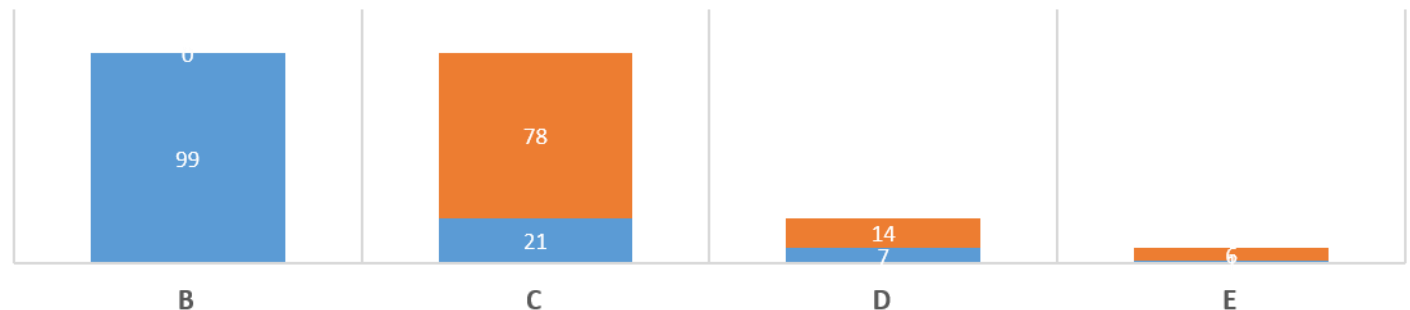
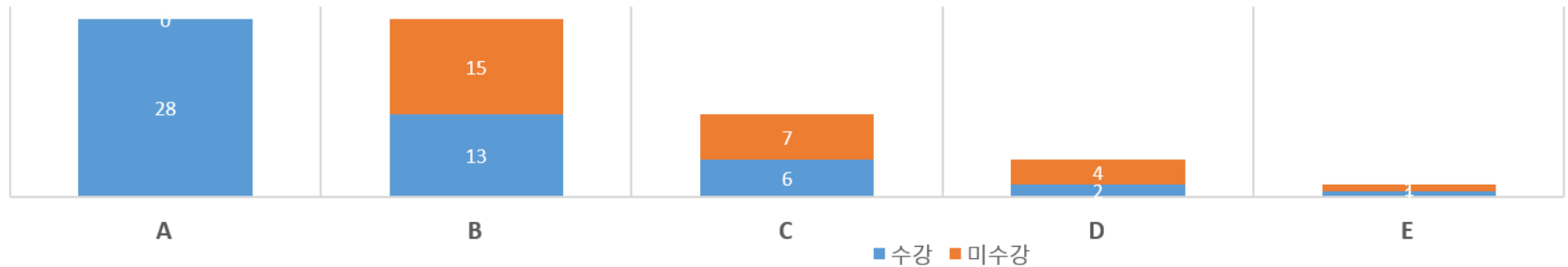
		지지도(support)				
		0.002	0.005	0.01	0.015	0.02
신뢰도(confidence)	0.1	2,264	196	40	19	6
	0.2	1,363	106	24	12	4
	0.3	928	40	4	2	1
	0.4	688	16	3	1	1
	0.5	592	13	2	1	1

# 사례 4: 기업체 교육프로그램 추천

## ❖ 생성 규칙 분석 I: Chain 규칙 발굴을 통한 교육프로그램 단계적 추천



■ 수강 ■ 미수강

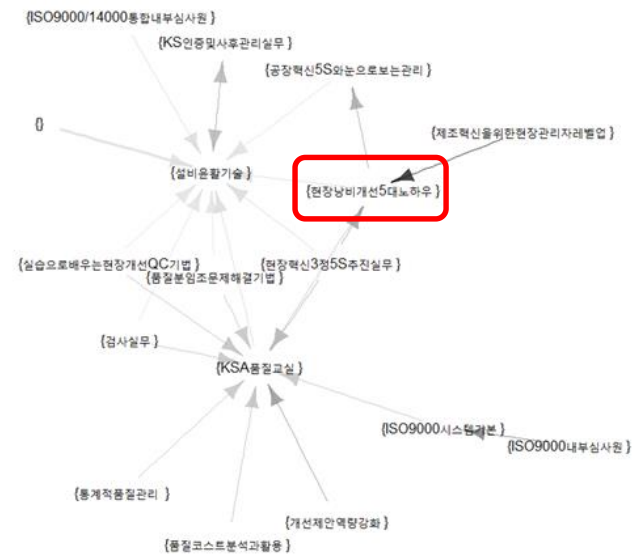


# 사례 4: 기업체 교육프로그램 추천

## ❖ 생성 규칙 분석 2: 수강기업수에 비해 규칙 등장 횟수가 높은 교육프로그램 발굴

순위	교육명	수강 기업수	순위	교육명	규칙등장 횟수
1	설비유효기술	1211	1	설비유효기술	11
2	KS인증및사후관리실무	234	2	KSA품질교실	10
3	KSA품질교실	201	3	현장낭비개선5대노하우	5
4	공장혁신5S와눈으로보는관리	148	4	현장혁신3정5S추진실무	3
5	ISO9000/14000통합내부심사원	122	5	품질분임조문제해결기법	2
6	ISO9000시스템기본	117	6	실습으로배우는현장개선QC기법	2
7	실습으로배우는현장개선QC기법	107	7	공장혁신5S와눈으로보는관리	2
8	품질분임조문제해결기법	103	8	검사실무	2
9	현장낭비개선5대노하우	99	9	KS인증및사후관리실무	2
10	검사실무	93	10	ISO9000시스템기본	2

Graph for 24 rules



# 사례 5: 한의학 처방전 분석

## ❖ 방약합편 분석

- 한의학 고문서에 나타난 텍스트를 분석하여 증상(symptoms)과 처방약재(medicines) 간의 관계를 규명 (Yang et al., 2013)

연구대상 : 방약합편 (方藥合編)

- 한의학의 경험과 지식 축적은 주로 서적의 형태로 정리
- 방약합편 : 동의보감이 발간된 이후, 병증을 중심으로 두고 각 병증의 말미에 처방을 제시한 서적
- 병증에 따라 유용한 처방을 한눈에 제시
- 처방에 필요한 약재(본초)들을 명시



Materia medica/prescription	Ginseng Radix	Citri Pericarpium	Pinelliae Rhizoma	Poria(red)	...	...
Prescription 1	1	0	0	0		
Prescription 2	1	1	0	1		
Prescription 3	0	0	0	1		
Prescription 4	0	0	1	1		
...						
Prescription 521	1	1	0	0		



# 사례 5: 한의학 처방전 분석

## ❖ 방약합편 분석

- 한의학 고문서에 나타난 텍스트를 분석하여 증상(symptoms)과 처방 약재(medicines) 간의 관계를 규명 (Yang et al., 2013)

Antecedent (symptom)	Consequent (herbal materials)	Support (%)	Confidence (%)	Lift
Coughs	Pinelliae Rhizoma	4.2	46.0	2.2
	Citri Pericarpium	3.9	42.9	1.3
	Rehmanniae Radix Preparat	2.2	23.8	1.8
	Poria(red)	2.0	22.2	1.3
	Poria(white)	2.0	22.2	1.1
	Armeniacae Semen	1.9	20.6	3.0
Overexertion/Fatigue	Rehmanniae Radix Preparat	3.5	70.5	4.5
	Angelicae Gigantis Radix	3.3	67.6	1.9
	Ginseng Radix	2.0	41.1	1.4
	Poria(white)	1.9	38.2	1.9
	Dioscoreae Rhizoma	1.7	35.3	5.0
	Paeoniae Radix Alba	1.7	35.3	1.8
	Atractylodis Rhizoma Alba	1.6	32.3	1.2
	Corni Fructus	1.4	29.4	5.8
	Capreoli Cornu	1.3	26.5	8.2
	Astragali Radix	1.3	26.5	2.4
	Achyranthis Radix	1.2	23.5	5.0
	Pulvis Aconiti Tuberis Purificatum	1.2	23.5	2.8
	Cinnamomi Cortex Spissus	1.2	23.5	1.8
	Moutan Cortex	1.0	20.6	3.6
	Schizandrae Fructus	1.0	20.6	3.4



# Table of Contents

---

I

Association Rules: Theory

II

Association Rules: Applications

III

R Exercise

# Association Rules

## ❖ Package “arules” & “arulesViz”

### Package ‘arules’

February 21, 2014

**Version** 1.1-2

**Date** 2014-2-21

**Title** Mining Association Rules and Frequent Itemsets

**Description** Provides the infrastructure for representing, manipulating and analyzing transaction data and patterns (frequent itemsets and association rules). Also provides interfaces to C implementations of the association mining algorithms Apriori and Eclat by C. Borgelt.

**Classification/ACM** G.4, H.2.8, I.5.1

**URL** <http://R-Forge.R-project.org/projects/arules/>,  
<http://lyle.smu.edu/IDA/arules/>

**Depends** R (>= 2.14.2), Matrix (>= 1.0-0)

**Imports** stats, methods

**Suggests** pmml, arulesViz, testthat

### Package ‘arulesViz’

March 11, 2014

**Version** 0.1-9

**Date** 2014-03-10

**Title** Visualizing Association Rules and Frequent Itemsets

**Author** Michael Hahsler and Sudheer Chelluboina

**Maintainer** Michael Hahsler <mhahsler@lyle.smu.edu>

**Depends** R (>= 2.14.0), arules (>= 1.0-5), grid

**Imports** scatterplot3d, vcd, seriation, igraph

**Suggests** iplots, Rgraphviz, graph

**Description** Various visualization techniques for association rules and itemsets. The packages also includes several interactive visualizations for rule exploration. This package extends package arules.

# Association Rules: Data Frame

## ❖ Titanic Data

```

1 # Association Rules -----
2 # arules and arulesviz packages install
3 install.packages("arules", dependencies = TRUE)
4 install.packages("arulesviz", dependencies = TRUE)
5
6 library(arules)
7 library(arulesviz)
8 library(wordcloud)
9
10 # Load titanic data set
11 titanic <- read.delim("titanic.txt", dec=",")
12 str(titanic)
13 head(titanic)

> # Load titanic data set
> titanic <- read.delim("titanic.txt", dec=",")
> str(titanic)
'data.frame': 1313 obs. of  5 variables:
 $ Name      : Factor w/ 1310 levels "Abbing, Mr Anthony",...: 22 25 26 27 24 31 45 46 50 54 ...
 $ Pclass    : Factor w/ 3 levels "1st","2nd","3rd": 1 1 1 1 1 1 1 1 1 1 ...
 $ Age       : Factor w/ 75 levels "0.17","0.33",...: 28 18 30 24 5 48 66 39 60 73 ...
 $ Sex       : Factor w/ 2 levels "female","male": 1 1 2 1 2 2 1 2 1 2 ...
 $ Survived: int  1 0 0 0 1 1 1 0 1 0 ...
> head(titanic)

```

	Name	Pclass	Age	Sex	Survived
1	Allen, Miss Elisabeth Walton	1st	29	female	1
2	Allison, Miss Helen Loraine	1st	2	female	0
3	Allison, Mr Hudson Joshua Creighton	1st	30	male	0
4	Allison, Mrs Hudson JC (Bessie Waldo Daniels)	1st	25	female	0
5	Allison, Master Hudson Trevor	1st	0.92	male	1
6	Anderson, Mr Harry	1st	47	male	1

# Association Rules: Data Frame

## ❖ Data Preprocessing

- Categorize a numeric variable, remove NA, etc.

```

15 # Remove "Name" column and group "Age" column
16 titanic_ar <- titanic[,2:5]
17 titanic_ar$Age = as.character(titanic_ar$Age)
18 c_idx <- which(as.numeric(titanic_ar$Age) < 20)
19 a_idx <- which(as.numeric(titanic_ar$Age) >= 20)
20 na_idx <- which(is.na(titanic_ar$Age))
21
22 titanic_ar$Age[c_idx] <- "Child"
23 titanic_ar$Age[a_idx] <- "Adult"
24 titanic_ar$Age[na_idx] <- "Unknown"
25
26 # Convert the attributes to factor
27 titanic_ar$Age <- as.factor(titanic_ar$Age)
28 titanic_ar$Survived <- as.factor(titanic_ar$Survived)

```

	PClass	Age	Sex	Survived
1	1st	Adult	female	1
2	1st	Child	female	0
3	1st	Adult	male	0
4	1st	Adult	female	0
5	1st	Child	male	1
6	1st	Adult	male	1
7	1st	Adult	female	1
8	1st	Adult	male	0
9	1st	Adult	female	1
10	1st	Adult	male	0

# Association Rules: Data Frame

## ❖ Find rules (default setting)

```
30 # Rule generation by Apriori algorithm with default settings
31 rules <- apriori(titanic_ar)
32 inspect(rules)
```

```
> rules <- apriori(titanic_ar)
```

parameter specification:

confidence	minval	smax	arem	aval	originalsupport	support	minlen	maxlen	target	ext
0.8	0.1	1	none	FALSE	TRUE	0.1	1	10	rules	FALSE

algorithmic control:

filter	tree	heap	memopt	load	sort	verbose
0.1	TRUE	TRUE	FALSE	TRUE	2	TRUE

```
apriori - find association rules with the apriori algorithm
version 4.21 (2004.05.09)      (c) 1996-2004  Christian Borgelt
set item appearances ... [0 item(s)] done [0.00s].
set transactions ... [10 item(s), 1313 transaction(s)] done [0.00s].
sorting and recoding items ... [10 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 done [0.00s].
writing ... [16 rule(s)] done [0.00s].
creating s4 object ... done [0.00s].
```

# Association Rules: Data Frame

## ❖ Find rules (default setting)

```
> inspect(rules)
```

	lhs	rhs	support	confidence	lift
1	{PClass=3rd}	=> {Survived=0}	0.4364052	0.8059072	1.226137
2	{Sex=male}	=> {Survived=0}	0.5399848	0.8331375	1.267566
3	{Survived=0}	=> {Sex=male}	0.5399848	0.8215527	1.267566
4	{PClass=2nd, Sex=male}	=> {Survived=0}	0.1127190	0.8554913	1.301576
5	{PClass=2nd, Survived=0}	=> {Sex=male}	0.1127190	0.9192547	1.418309
6	{PClass=1st, Sex=female}	=> {Survived=1}	0.1020564	0.9370629	2.734141
7	{Sex=female, Survived=0}	=> {PClass=3rd}	0.1005331	0.8571429	1.582881
8	{PClass=3rd, Age=Unknown}	=> {Survived=0}	0.2536177	0.8473282	1.289156
9	{Age=Unknown, Sex=male}	=> {Survived=0}	0.2566641	0.8798956	1.338706
10	{Age=Unknown, Survived=0}	=> {Sex=male}	0.2566641	0.8023810	1.237986
11	{Age=Adult, Sex=male}	=> {Survived=0}	0.2482864	0.8253165	1.255667
12	{Age=Adult, Survived=0}	=> {Sex=male}	0.2482864	0.8693333	1.341286
13	{PClass=3rd, Sex=male}	=> {Survived=0}	0.3358720	0.8837675	1.344596
14	{PClass=3rd, Age=Unknown, Sex=male}	=> {Survived=0}	0.1957350	0.9081272	1.381658
15	{PClass=3rd, Age=Adult, Sex=male}	=> {Survived=0}	0.1142422	0.8670520	1.319165
16	{PClass=3rd, Age=Adult, Survived=0}	=> {Sex=male}	0.1142422	0.8064516	1.244267

# Association Rules: Data Frame

## ❖ Find rules (customized setting)

```

34 # Rule generation by Apriori algorithm with custom settings
35 rules <- apriori(titanic_ar, parameter = list(minlen = 3, support = 0.1, conf = 0.8),
36               appearance = list(rhs = c("Survived=0", "Survived=1"), default="lhs"))
37 inspect(rules)
38
39 # Plot the rules
40 plot(rules, method="scatterplot")
41 plot(rules, method="graph", control=list(type = "items", alpha = 1))
42 plot(rules, method="paracoord", control=list(reorder=TRUE))

```

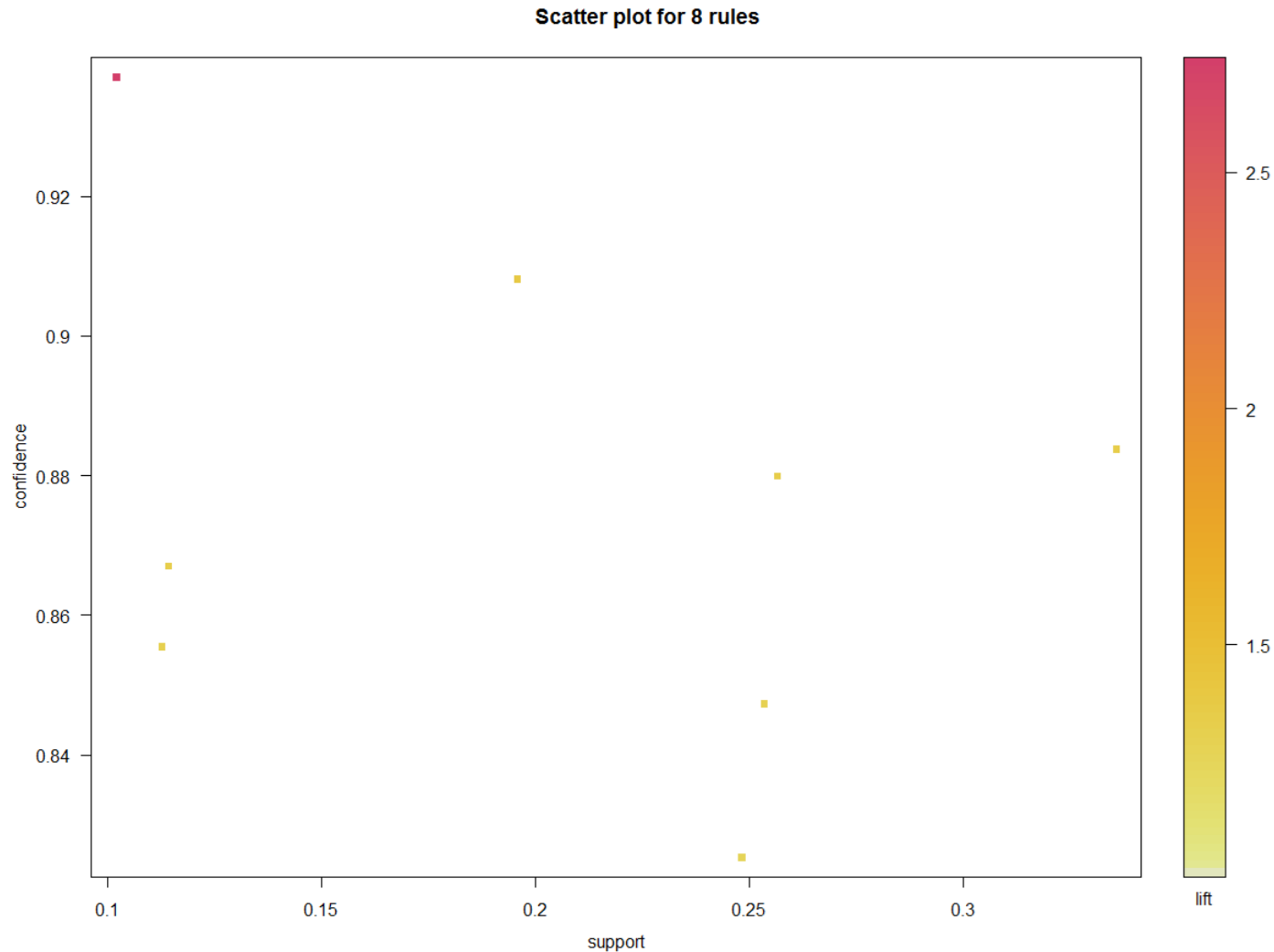
```

> inspect(rules)
lhs                rhs                support confidence    lift
1 {Pclass=2nd,      => {Survived=0} 0.1127190  0.8554913  1.301576
  Sex=male}
2 {Pclass=1st,      => {Survived=1} 0.1020564  0.9370629  2.734141
  Sex=female}
3 {Pclass=3rd,      => {Survived=0} 0.2536177  0.8473282  1.289156
  Age=Unknown}
4 {Age=Unknown,     => {Survived=0} 0.2566641  0.8798956  1.338706
  Sex=male}
5 {Age=Adult,       => {Survived=0} 0.2482864  0.8253165  1.255667
  Sex=male}
6 {Pclass=3rd,      => {Survived=0} 0.3358720  0.8837675  1.344596
  Sex=male}
7 {Pclass=3rd,      => {Survived=0} 0.1957350  0.9081272  1.381658
  Age=Unknown,
  Sex=male}
8 {Pclass=3rd,      => {Survived=0} 0.1142422  0.8670520  1.319165
  Age=Adult,
  Sex=male}

```

# Association Rules: Data Frame

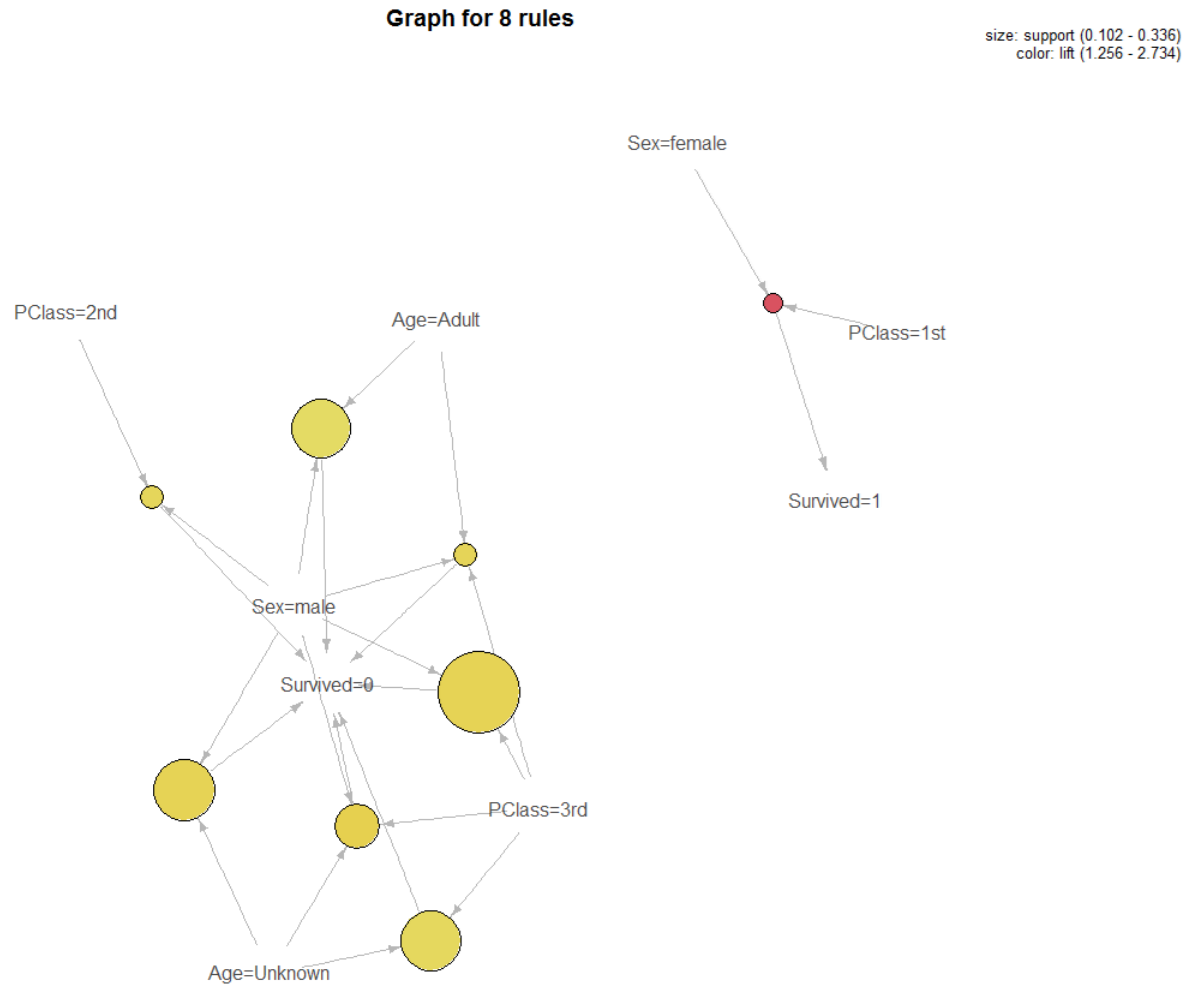
## ❖ Visualize the rules





# Association Rules: Data Frame

## ❖ Visualize the rules



### Parallel coordinates plot for 8 rules



# Association Rules: Transaction Data

## ❖ Groceries shopping data

```
44 # Load transaction data "Groceries"
45 data("Groceries")
46 summary(Groceries)
47 str(Groceries)
48 inspect(Groceries)
```

```
> summary(Groceries)
```

```
transactions as itemMatrix in sparse format with
9835 rows (elements/itemsets/transactions) and
169 columns (items) and a density of 0.02609146
```

```
most frequent items:
```

whole milk	other vegetables	rolls/buns	soda	yogurt	(other)
2513	1903	1809	1715	1372	34055

```
element (itemset/transaction) length distribution:
sizes
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
2159	1643	1299	1005	855	645	545	438	350	246	182	117	78	77	55	46	29	14	14	9	11	4	6	1
26	27	28	29	32																			
1	1	1	3	1																			

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	2.000	3.000	4.409	6.000	32.000

```
includes extended item information - examples:
```

	labels	level2	level1
1	frankfurter	sausage	meet and sausage
2	sausage	sausage	meet and sausage
3	liver loaf	sausage	meet and sausage

# Association Rules: Transaction Data

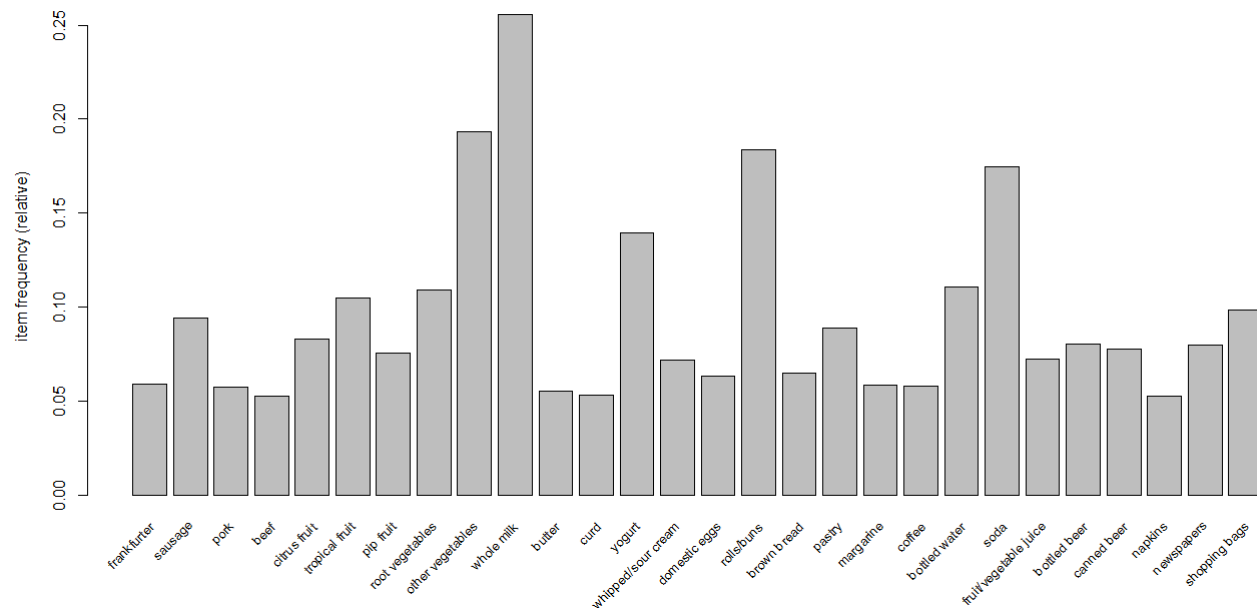
## ❖ Groceries shopping data

### ■ Item inspection

```

50 # Item inspection
51 itemName <- itemLabels(Groceries)
52 itemCount <- itemFrequency(Groceries)*9835
53
54 col <- brewer.pal(8, "Dark2")
55 wordcloud(words = itemName, freq = itemCount, min.freq = 1, scale = c(10, 0.2), col = col , random.order = FALSE)
56
57 itemFrequencyPlot(Groceries, support = 0.05, cex.names=0.8)

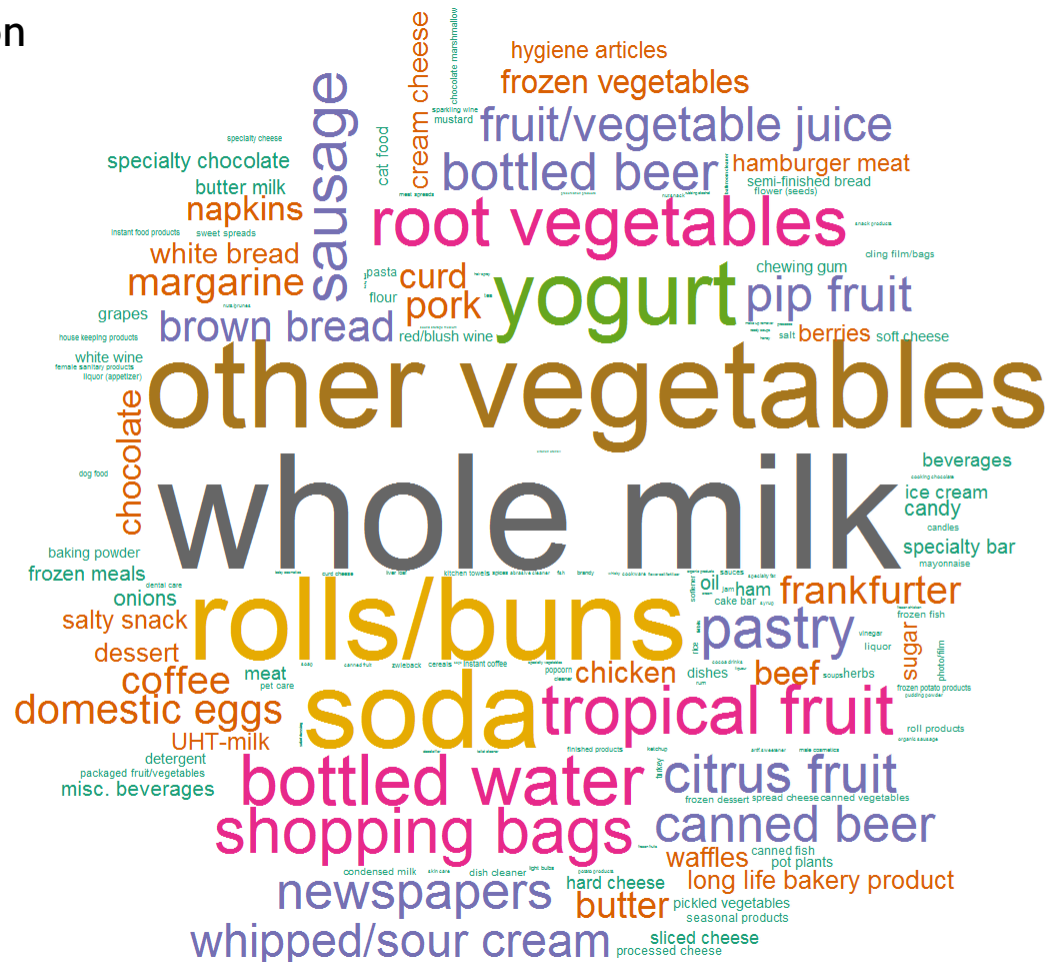
```



# Association Rules: Transaction Data

## ❖ Groceries shopping data

### ■ Item inspection



# Association Rules: Transaction Data

## ❖ Find and visualize rules

```

59 # Rule generation by Apriori
60 rules <- apriori(Groceries, parameter=list(support=0.001, confidence=0.5))
61 rules
62
63 # List the first three rules with the highest lift values
64 inspect(head(sort(rules, by="lift"),3))
65
66 # Save the rules in a text file
67 write.csv(as(rules, "data.frame"), "Groceries_rules.csv", row.names = FALSE)
68
69 # Plot the rules
70 plot(rules)
71 plot(rules, method="grouped")

```

```

> inspect(head(sort(rules, by="lift"),3))

```

	lhs	rhs	support	confidence	lift
1	{Instant food products, soda}	=> {hamburger meat}	0.001220132	0.6315789	18.99565
2	{soda, popcorn}	=> {salty snack}	0.001220132	0.6315789	16.69779
3	{flour, baking powder}	=> {sugar}	0.001016777	0.5555556	16.40807

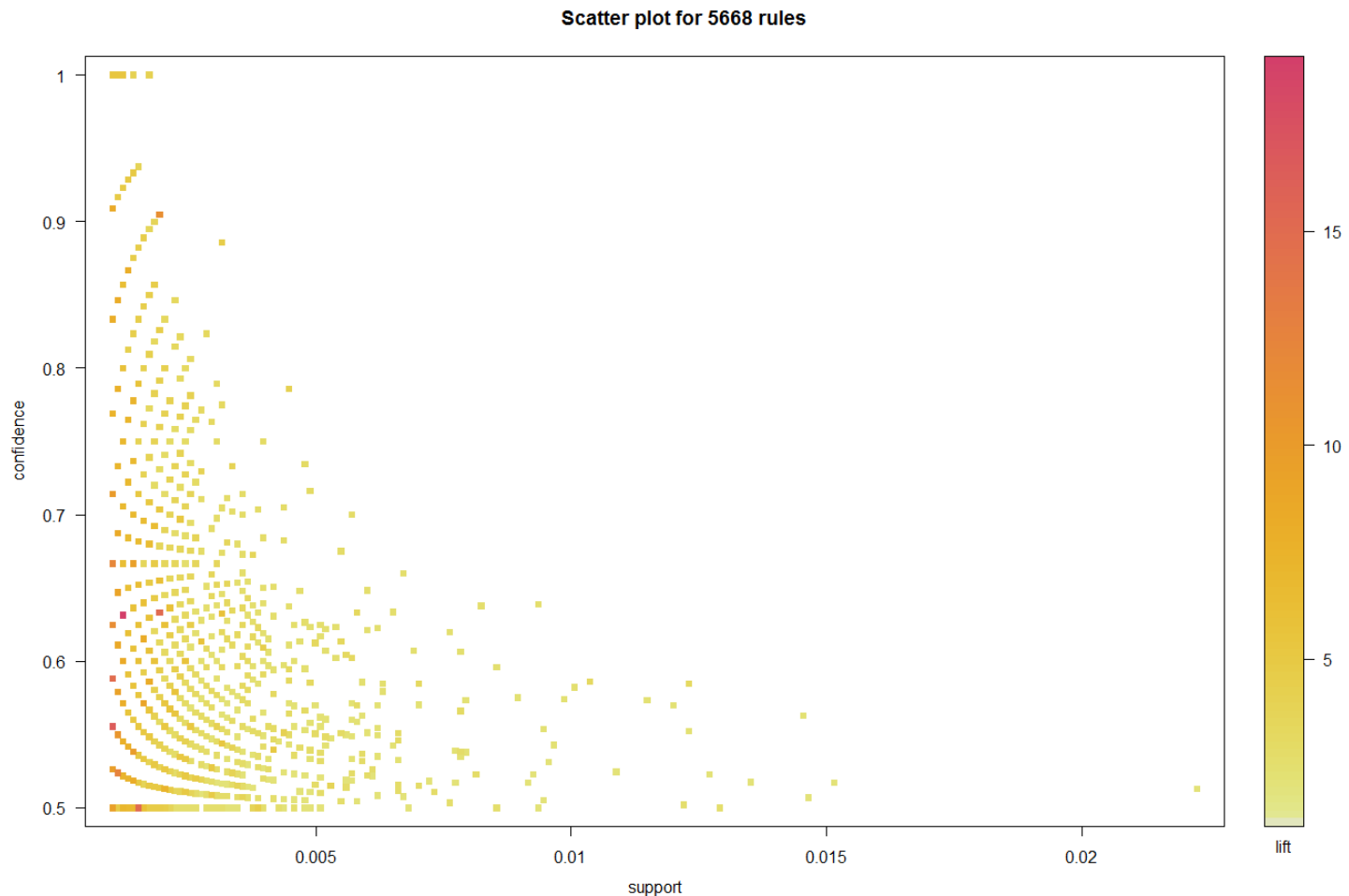
# Association Rules: Transaction Data

## ❖ Find and visualize rules

	A	B	C	D
1	rules	support	confidence	lift
2	{honey} => {whole milk}	0.001118	0.733333	2.870009
3	{tidbits} => {rolls/buns}	0.00122	0.521739	2.836542
4	{cocoa drinks} => {whole milk}	0.001322	0.590909	2.312611
5	{pudding powder} => {whole milk}	0.001322	0.565217	2.212062
6	{cooking chocolate} => {whole milk}	0.001322	0.52	2.035097
7	{cereals} => {whole milk}	0.00366	0.642857	2.515917
8	{jam} => {whole milk}	0.002949	0.54717	2.141431
9	{specialty cheese} => {other vegetables}	0.00427	0.5	2.584078
10	{rice} => {other vegetables}	0.003965	0.52	2.687441
11	{rice} => {whole milk}	0.004677	0.613333	2.400371
12	{baking powder} => {whole milk}	0.009253	0.522989	2.046793
13	{liver loaf,yogurt} => {whole milk}	0.001017	0.666667	2.609099
14	{tropical fruit,curd cheese} => {other vegetables}	0.001017	0.666667	3.445437
15	{curd cheese,rolls/buns} => {whole milk}	0.001017	0.625	2.446031
16	{other vegetables,curd cheese} => {whole milk}	0.00122	0.571429	2.236371
17	{whole milk,curd cheese} => {other vegetables}	0.00122	0.521739	2.696429
18	{other vegetables,cleaner} => {whole milk}	0.001017	0.625	2.446031
19	{liquor,red/blush wine} => {bottled beer}	0.001932	0.904762	11.23527
20	{soda,liquor} => {bottled beer}	0.00122	0.571429	7.09596

# Association Rules: Transaction Data

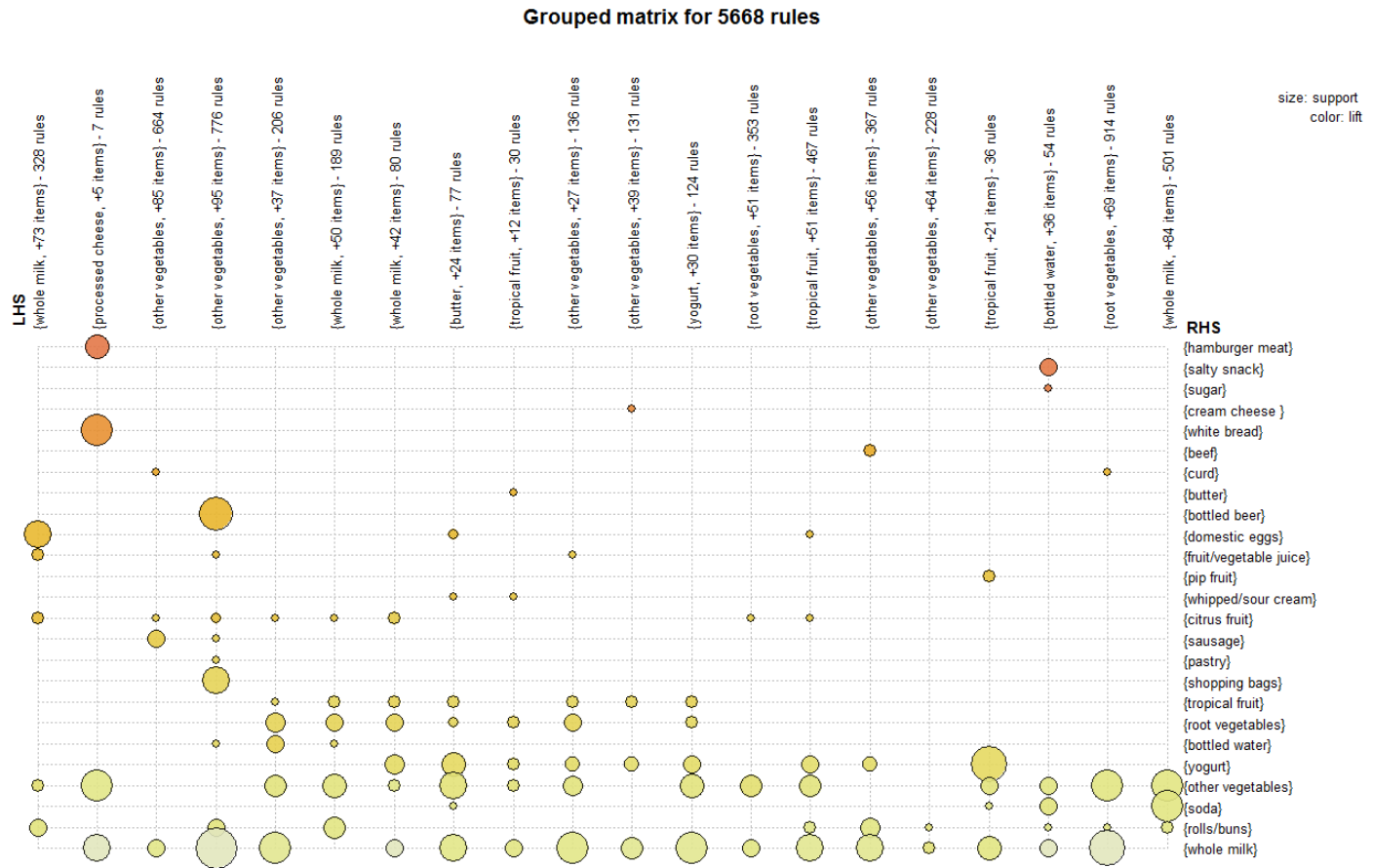
## ❖ Find and visualize rules





# Association Rules: Transaction Data

## ❖ Find and visualize rules



# Q & A

